Data Literacy for the Language Sciences

A very gentle introduction to statistics and data visualisation in R

Elen Le Foll

2024-04-17

Table of contents

Preface

> **Warning**
>
> This textbook draft is very much **work in progress**. I intend to progressively add to it over the course of the summer semester 2024.
> This first draft is intended as complementary materials to my summer semester M.A. class: "More than counting words: Introduction to statistics and data visualisation for linguists" taught at the University of Cologne.
> Student feedback on this first draft is very welcome!

Who is this book for?

This textbook is intended as a very gentle introduction to basic principles of data management, statistics, and data visualisation using the programming language and environment 'R'. The target audience are students and researchers in the language sciences, including (applied) linguistics, language teaching, and language education research. The rationale for this textbook is based on my personal observations, in teaching and consulting both students and researcher colleagues, that many so-called 'introductory' textbooks assume previous knowledge and skills that all have or go through contents at too fast a pace for many humanities scholars who often come with little to no experience with programming and/or statistics.

The aim of this textbook is by no means to replace any of the brilliant existing textbooks aimed at imparting statistical literacy for linguistics research, but rather to provide a stepping stone towards being able to make the most of these wonderful existing resources. A (work-in-progress) list of next-step resources is included in Appendix A.

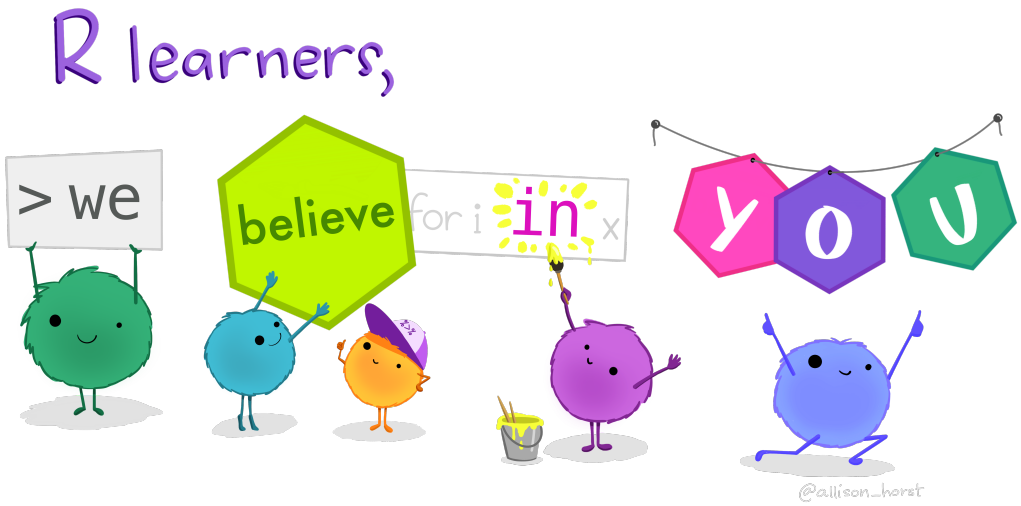

Figure 1: *Artwork by @allison_horst*

1 Open Scholarship

This book aims to provide a stepping stone for students and scholars of traditionally less quantitative and computational disciplines (such as some branches of linguistics and language education research) to gather first (hopefully positive!) experiences with statistical and computational approaches to working with empirical data[1]. The underlying belief is that these methods ought to be accessible to all, regardless of their academic background or personal circumstances. To this end, this book embraces the principles of Open Scholarship.

Open Scholarship "reflects the idea that knowledge of all kinds should be openly shared, transparent, rigorous, reproducible, replicable, accumulative, and inclusive (allowing for all knowledge systems)" (Parsons et al. 2022). For this to be the case, teaching materials need to shared openly and the tools and software taught in these resources need to be freely accessible, too. In the following, we will briefly consider the role of Open Educational Resources (OERs) and open-source software in our pursuit of Open Scholarship.

1.1 Open Source

In line with its aim to provide an accessible introduction to statistics and data visualisation, this textbook relies exclusively on open-source software and programming languages, foremost `LibreOffice Calc`, `R` and `RStudio`. Open source refers to software whose source code is available under a license that grants anyone the rights to study, modify, and distribute the software to anyone and for any purpose. If we think of a software application as a cake, the source code is like its recipe. It contains the list of ingredients and the steps to bake the cake. Open source means that the recipe is publicly available. You can access it, read it, and use it to bake the cake. You can also modify it to add your own twist, such as adding a new ingredient or making it vegan, and share it with others. In the context of software, this allows many people to collaborate, make improvements, and share their versions, resulting in better and more diverse software.

Using open-source software in this introductory textbook means that anyone[2] can download, install and use the required software at no cost. However, it is very important to note that not all free software (*freeware*) is open source. Let us illustrate the difference by comparing

[1] Emprical data is based on what is experienced or observed rather than on theory alone.
[2] Provided that they have access to the internet and a functioning personal computer.

different spreadsheet programmes as, in the following chapter, we will begin exploring tabular data structures in a spreadsheet programme.

The most most widely used spreadsheet programme to date is undoubtedly `Microsoft Excel`. Excel is a commercial, proprietary spreadsheet editor which forms part of the Microsoft 365 package. As such, to use Excel on your personal computer, you need to buy a license or be a member of an organisation (e.g., your university or company) that pays for such a license. It is true that Microsoft now also offers a free (functionally limited) web-based version of Excel, yet this still does not make it open source. This is because Microsoft does not share the source code of any Excel version, which means that, even if they are giving away free cake, we do not have the recipe to bake the cake ourselves should the company decide to start charging money for the cake or to no longer distribute it at all! Similarly, you may be familiar with a popular, web-based alternative to Excel: `Google Sheets`. Whilst it is (currently) free to use, as the name suggests, Google Sheets is owned by Google and is therefore not open source, either. By contrast, `LibreOffice Calc` is a project of The Document Foundation (TDF) that provides a popular, free, open-source office productivity software suite comparable to Microsoft 365 called `LibreOffice`. LibreOffice is developed collaboratively by very many different people across the world who all do so on a volunteer basis. The Document Foundation estimates that there are 200 million active LibreOffice users worldwide, about 25% of whom are thought to be students (figures from 2018, see LibreOffice 2024). Its popularity is likely due to the fact that it not only uses open formats (e.g., `.odt` and `.ods`), but can also open and save to a range of popular formats including those used by Microsoft (e.g., `.docx` and `.xlsx`).

Quiz time!

1) Which of these is an open-source alternative to Microsoft Word?

2) Which of these is an open-source alternative to Microsoft Powerpoint?

3) Not only can software be open source, programming languages can, too. In fact, most modern programming languages are open source. In this book, we will focus on the open-source programming language `R`. Which of these is another open-source programming language?

4) There are also many open-source operating systems. Which of these is an open-source alternative to the operating system Windows?

Task 1

Your first task is to **download** and **install LibreOffice** as we will use its spreadsheet editor, **LibreOffice Calc**, in the next few chapters.

- LibreOffice is available for Windows, Mac and Linux. You can download it from here: https://www.libreoffice.org/download/download-libreoffice/.
- You will find detailed installation instructions here: https://www.libreoffice.org/get-help/install-howto/.
- Detailed documentation is also available in many different languages: https://documentation.libreoffice.org/en/english-documentation/

1.2 Open Education

The web-based version of this textbook is published as an Open Educational Resource (OER) under the Creative Commons license: CC BY-NC-SA. This means that it is free to read and use, as well as edit, remix, and expand upon, provided that a) the original author and source is mentioned (as indicated by BY), b) any derived version is not made into a commercial product (NC stands for non-commercial), and c) that any derived versions of this textbook (e.g., a translated version or a version adapted for historians) are also shared with this same license (SA stands for share alike).

In line with the principles of Open Education, all of the datasets that we will work with in this textbook have been published in Open Access, which means that we can freely use them to learn about statistics and data visualisation using real datasets from published research studies in applied linguistics and language education.

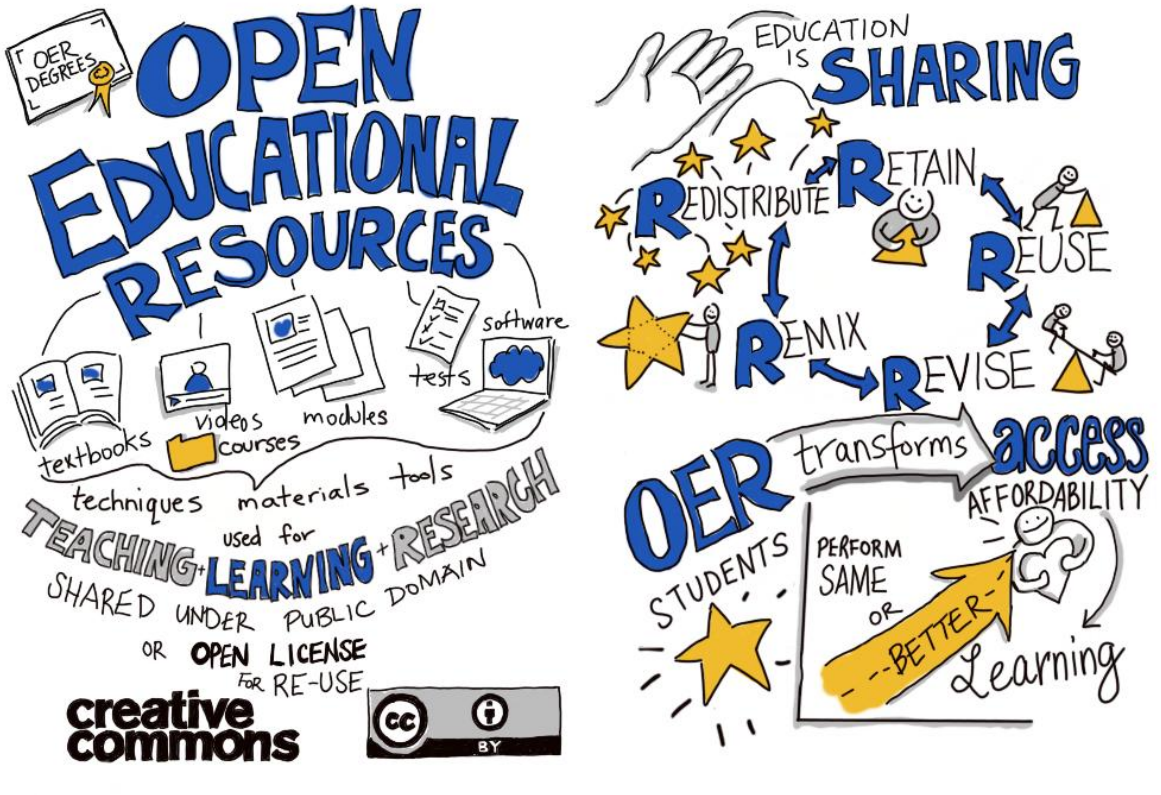

Figure 1.1: OER sketch note by Yvonne Stry

i Tips to go further

This chapter has simplified things considerably. To be considered open source, software distributions actually have to comply with ten criteria. You can read up on them here:

- https://opensource.org/osd

To find out more about the benefits of open-source software in the context of research, I recommend reading:

- https://book.the-turing-way.org/reproducible-research/open/open-source

To find out more about Open Educational Resources (OERs), I recommend exploring the following OER databases:

- https://oercommons.org/
- https://www.twillo.de/oer/web/

References

2024. LibreOffice. *Wikipedia*. https://en.wikipedia.org/w/index.php?title=LibreOffice&oldid=1218520104.

Parsons, Sam, Flávio Azevedo, Mahmoud M. Elsherif, Samuel Guay, Owen N. Shahim, Gisela H. Govaart, Emma Norris, et al. 2022. A community-sourced glossary of open scholarship terms. *Nature Human Behaviour*. Nature 6(3). 312–318. https://doi.org/10.1038/s41562-021-01269-4.

A Next-step resources

In the hope that this textbook has inspired you to dive deeper into the wonderful world of quantitative data analysis, statistics, data visualisation, and coding in R, here is a (work-in-progress) curated list of further resources to continue your learning journey! *Bon voyage!*

A.1 Recommended resources specific to the language sciences

- Brezina, Vaclav. 2018. Statistics in Corpus Linguistics: A Practical Guide. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316410899.
- Desagulier, Guillaume. 2017. Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics (Quantitative Methods in the Humanities and Social Sciences). Cham: Springer International Publishing.
- Gries, Stefan Thomas. 2013. Statistics for linguistics with R: a practical introduction. 2nd revised edition. Berlin: De Gruyter Mouton.
- LADAL contributors. Tutorials of the Language Technology and Data Analysis Laboratory. https://ladal.edu.au/tutorials.html `Open Educational Resource`.
- Levshina, Natalia. 2015. How to do linguistics with R: Data exploration and statistical analysis. Amsterdam: John Benjamins.
- Schneider, Dr Gerold & Max Lauber. 2020. Statistics for Linguists. https://dlf.uzh.ch/openbooks/statisticsforlinguists/ `Open Educational Resource`.
- Winter, Bodo. 2019. Statistics for Linguists: An Introduction Using R. New York: Routledge. https://doi.org/10.4324/9781315165547.

A.2 Further Open Educational Resources (in no particular order)

- Diez, David, Mine Cetinkaya-Rundel, Christopher Barr & OpenIntro. 2015. OpenIntro Statistics. Leanpub. https://leanpub.next/os.
- Guide to Effect Sizes and Confidence Intervals: https://matthewbjane.quarto.pub/guide-to-effect-sizes-and-confidence-intervals/

- Happy Git and GitHub for the useR: https://happygitwithr.com/
- Quarto & reproducibility: https://ucsbcarpentry.github.io/Reproducible-Publications-with-RStudio-Quarto/index.html
- Modern Data Visualization with R: https://rkabacoff.github.io/datavis
- Building reproducible analytical pipelines with R: https://raps-with-r.dev/
- Modern Plain Text Computing: https://mptc.io/content/01-content.html
- https://www.data-to-viz.com/
- Interpreting data visualisation: https://pressbooks.library.torontomu.ca/criticaldataliteracy/
- Improve your statistical inferences: https://lakens.github.io/statistical_inferences/
- What they forgot to teach you about R: https://rstats.wtf/
- Introduction to Data Science: https://florian-huber.github.io/data_science_course/book/cover.html
- Data Science in Education Using R: https://datascienceineducation.com/
- Models Demystified: A Practical Guide from t-tests to Deep Learning https://m-clark.github.io/book-of-models/
- Data Visualization in R https://datavizf23.classes.andrewheiss.com/